CLAIMS

What is claimed is:

1. A disk mirroring apparatus comprising two or more processing assemblies each processing assembly further comprising:

a processing unit, further comprising one or more general purpose processors, a memory, one or more disk controllers and a network interface; and

a disk coupled to the processing unit, the storage of the disk being logically divided into at least two data segments, wherein a first data segment of a first disk in a first processing assembly is mirrored by a secondary data segment of a second disk in a second processing assembly.

2. The disk mirroring apparatus of claim 1 in which the processing assemblies further comprise additional disks coupled to the processing unit of the processing assembly, each disk logically divided into at least two segments, wherein a first data segment of each of the plurality of additional disks in a first processing assembly is mirrored by a secondary data segment of one of the plurality of additional disks in a second processing assembly.

3. The disk mirroring apparatus of claim 1 in which the disk is logically divided into at least three data segments, wherein data in a third data segment is not mirrored.

4. The disk mirroring apparatus of claim 3 in which a choice can be made about whether a new data item is to be mirrored or not mirrored by specifying whether the new data item is to be stored in a first data segment or a third data segment.

5. The disk mirroring apparatus of claim 1 further comprising:

a plurality of host computers that request modifications to data stored on the disk of a processing assembly by communicating with the processing unit of the processing assembly via its network interface.

6. The disk mirroring apparatus of claim 5 wherein the modifications requested by one of the host computers to data on the disk of a first processing assembly are also automatically performed on the disk of a second processing assembly, without intervention from the host computer.

7. The disk mirroring apparatus of claim 6 wherein the processing unit of the first processing assembly stores the data requested by the host computer in the first data segment of its disk and forwards the data to the processing unit of the second processing assembly to mirror the data on the secondary data segment of its disk.

8. The disk mirroring apparatus of claim 7 wherein the processing unit of the second processing assembly receives the data that was forwarded by the processing unit of the first processing assembly, and writes that data to the secondary data segment of its disk.

9. The disk mirroring apparatus of claim 1, wherein the processing unit further comprises:

a mirror manager that manages mirroring of a block of data in the first segment of its disk into the secondary data segment of the disk of a second processing assembly.

10. The disk mirroring apparatus of claim 9, wherein the mirror manager manages mirroring of one or more database records in the first data segment of its disk into the secondary data segment of the disk of a the second processing assembly.

11. The disk mirroring apparatus of claim 9 wherein the mirror manager operates autonomously from any host computer.

12. The disk mirroring apparatus of claim 7, wherein the processing unit of the second processing assembly may defer writing data to the secondary data segment of its disk until it receives a commit command.

13. The disk mirroring apparatus of claim 1, wherein a processing unit further comprises:

a storage manager, wherein the storage manager assigns logical blocks which map to disk tracks having the fastest data transfer time to the first data segment of its disk and also assigns logical blocks which map to disk tracks having a slower data transfer time to the secondary data segment of its disk.

14. The disk mirroring apparatus of claim 13 wherein the assignment of blocks to segments is made once by the storage manager when the processing assembly is first made available for use.

15. The disk mirroring apparatus of claim 13 wherein the assignment of blocks to segments made by the storage manager may occur dynamically in response to data storage requests, such that blocks are allocated in the first data segment in response to a request to store data in the first data segment and blocks are allocated in the secondary data segment in response to a request to store a mirror copy of the data in the secondary data segment.

16. The disk mirroring apparatus of claim 1 wherein the secondary data segment of the disk of the second processing assembly is a logical mirror of the first data segment of the disk or the first processing assembly.

17. The disk mirroring apparatus of claim 1 further comprising:

a system manager, wherein the system manager controls the distribution map so as to evenly distribute data between the disks.

18. The disk mirroring apparatus of claim 17 wherein the system manager runs on a plurality of host computers.

19. The disk mirroring apparatus of claim 1 wherein a first segment of the disk of the second processing assembly is mirrored in a secondary data segment of the disk of the first processing assembly.

20. The disk mirroring apparatus of claim 1 wherein a first segment of the disk of the second processing assembly is mirrored in a secondary data segment of a disk of a third processing assembly.

21. The disk mirroring apparatus of claim 1 wherein the first data segment of the disk of the first processing assembly is mirrored in secondary data segments of the disks of two or more other processing assemblies.

22. The disk mirroring apparatus of claim 1 further comprising:

a spare processing assembly is activated by a system manager upon detecting failure of the disk or the processing unit of one of the plurality of processing assemblies.

23. The disk mirroring apparatus of claim 22 wherein

a second processing unit transmits to the processing unit of the spare processing assembly data stored in a secondary data segment of the disk coupled to the second processing unit, this secondary data segment having been a mirror of a first data segment of the disk of the failed processing assembly; and

the processing unit of the spare processing assembly stores the data it receives from the second processing unit in the first data segment of the disk of the spare processing assembly.

24. The disk mirroring apparatus of claim 22 wherein

a processing unit of a first processing assembly transmits to the processing unit of the spare processing assembly the data stored in a first data segment of the disk coupled to the first processing unit, this first data segment having been mirrored by the secondary data segment of the disk of the failed processing assembly; and

the processing unit of the spare processing assembly stores the data it receives from the first processing unit in the secondary data segment of the disk of the spare processing assembly.

25. The disk mirroring apparatus of claim 22 wherein the system manager creates a spare processing assembly by redistributing data stored on the disk of a first processing assembly among the disks of a subset of the plurality of processing assemblies, not including the first processing assembly, after which the first processing assembly can serve as the spare processing assembly.

26. The disk mirroring apparatus of claim 25 wherein the step of redistributing data from the first disk further comprises reassigning blocks in a distribution map.

27. The disk mirroring apparatus of claim 1 wherein the plurality of processing assemblies is subdivided into at least two sets.

28. The disk mirroring apparatus of claim 27 wherein first segments of disks in a first set of processing assemblies are mirrored in secondary data segments of disks in a second set of processing assemblies, and wherein first data segments

of the disks in the second set of processing assemblies are mirrored in secondary data segments of the disks in the first set of processing assemblies.

29. The disk mirroring apparatus of claim 28 wherein the probability of a double failure of both a processing assembly in the first set and a processing assembly in the second set within a given period of time is less than the probability of a double failure of two processing assemblies in the first set or the probability of a double failure of two processing assemblies in the second set within the same period of time.

30. The disk mirroring apparatus of claim 28 wherein the processing assemblies in the different sets are powered separately.

31. The disk mirroring apparatus of claim 28 wherein each set of processing assemblies is served by a separate network switch to which the network interfaces of the processing units of the processing assemblies in that set are coupled.

32. A method for disk mirroring in a system of multiple disks coupled to multiple processing units, said method comprising:

writing, by a first processing unit, first data to a first segment of a first disk coupled to the first processing unit;

forwarding, by the first processing unit, the first data to a second processing unit; and

writing, by the second processing unit, the first data to a secondary segment of a second disk coupled to the second processing unit.

33. A method for disk mirroring of claim 32 further comprising:

writing, by the second processing unit, second data to a first segment of the disk coupled to the second processing unit;

forwarding, by the second processing unit, the second data to an other processing unit; and

writing, by the other processing unit, the second data to a secondary segment of a disk coupled to the other processing unit.

34. A method for disk mirroring of claim 32 wherein mirroring is performed by the multiple processing units under direction of a mirror manager.

35. A method for disk mirroring of claim 32 further comprising:

assigning, by a storage manager, logical blocks which map to disk tracks having the fastest data transfer time to the first segment of the first disk; and

assigning, by the storage manager, logical blocks which map to disk tracks having a slower data transfer time to the secondary segment of the first disk.

36. A method for disk mirroring of claim 32 wherein the secondary segment of the second disk is a logical mirror of the first segment of the first disk.

37. A method for disk mirroring of claim 32 further comprising:

distributing, by a system manager, data between disks.

38. A method for disk mirroring of claim 37 wherein the step of distributing data comprises reassigning blocks in a distribution map.

39. A method for disk mirroring of claim 37 wherein the step of distributing data between disks is performed in case of a fail-over.

40. A method for disk mirroring of claim 32 further comprising:

forwarding, by the first processing unit, the first data to a third processing unit; and

writing, by the third processing unit, the first data to a secondary segment of a disk coupled to the third processing unit.

41. A method for disk mirroring of claim 32 further comprising:

rebuilding, in case of a processing unit failure, data on a disk associated with the failed processing unit using a spare processing unit.

42. A method for disk mirroring of claim 41 wherein the step of rebuilding further comprises:

rebuilding data stored on a first segment of the disk coupled to the failed processing unit, using a secondary data segment corresponding to the first data segment of the disk coupled to the failed processing unit.

43. A method for disk mirroring of claim 41 wherein the step of rebuilding further comprises rebuilding data stored on a secondary segment of the disk coupled to the failed processing unit using a primary data segment corresponding to the secondary data segment of the disk coupled to the failed processing unit.

44. A method for disk mirroring of claim 32 further comprising:

rebuilding, in case of a disk failure, data on a spare disk.

45. A method for disk mirroring of claim 44 wherein the step of rebuilding further comprises:

rebuilding data on a first segment of the spare disk using a secondary data segment corresponding to the first data segment of the failed disk.

46. A method for disk mirroring of claim 44 wherein the step of rebuilding further comprises:

rebuilding data on a secondary segment of the spare disk using a first data segment corresponding to the secondary data segment of the failed disk.

47. A method for disk mirroring of claim 44 further comprising:

creating a spare disk by redistributing data stored on the first disk among a subset of the plurality of disks.

48. A method for disk mirroring of claim 47 wherein the step of creating a spare disk further comprises reassigning blocks in a distribution map.

49. A method for disk mirroring of claim 32 further comprising:

subdividing the plurality of disks into at least two sets of disks, wherein first segments of disks in a first set of disks are mirrored in secondary segments of disks in a second set of disks, and wherein first segments of the disks in the second set of disks are mirrored in secondary segments of the disks in the first set of disks.

50. A method for disk mirroring of claim 32 further comprising:

rebuilding, in case of a sector failure on the first segment of the first disk, the failed sector using a corresponding sector on the secondary segment of the second disk.

51. A method for disk mirroring of claim 32 further comprising:

rebuilding, in case of a sector failure on the secondary sector of the second disk, the failed sector using a corresponding sector on the first segment of the first disk.

52. A method for disk mirroring of claim 32 further comprising:

modifying, by a system manager, mirroring topology in case of a topology-affecting event.

53. A method for disk mirroring of claim 52 wherein the topology-affecting event is at least one of: a disk failure and a processing unit failure.

54. A method for disk mirroring of claim 52 wherein the topology-affecting event is an addition of a new disk to the plurality of disks.

55. A method for disk mirroring of claim 52 wherein the topology-affecting event is a removal of one disk from the plurality of disks.

56. A method for disk mirroring of claim 52 wherein the topology-affecting event is partitioning of the plurality of disks into two or more sets of disks.

57. A method for fail-over processing in a system of multiple disks coupled to multiple processing units, where each of the multiple disks is logically divided into two or more segments, said method comprising:

maintaining, by a first and a second processing unit, a mirror of a first segment of a first disk in a secondary segment of a second disk;

swapping, in case of a failure of the first disk or the first processing unit, data in a distribution map pointing to the first segment of the first disk and the secondary segment of the second disk; and

responding, by the second processing unit, to commands directed to data stored on the first segment of the first disk, using data stored in the secondary segment of the second disk.

58. A method for fail-over processing of claim 57 wherein a command directed to data stored on first segments of the multiple disks are broadcast to the multiple processing units.

59. A method for fail-over processing of claim 58 further comprising:

responding twice, by the second processing unit, to the broadcasted command, one response involving the data stored on a first segment of the second disk and another response involving the data stored on the secondary segment of the second disk.

60. A method for regenerating a disk mirror in case of a disk failure in a system of multiple disks coupled to multiple processing units, said method comprising:

forwarding, by a first processing unit coupled to a disk containing a mirror of a first segment of the failed disk, data from a secondary segment of the coupled disk to a spare processing unit coupled to a spare disk;

writing, by the spare processing unit coupled to the spare disk, data received from the first processing unit to a first segment of the spare disk;

forwarding, by a second processing unit coupled to a disk containing a first segment mirrored by a secondary segment of the failed disk, data from the first segment of the coupled disk to the spare processing unit coupled to the spare disk; and

writing, by the spare processing unit coupled to the spare disk, data received from the second processing unit to a secondary segment of the spare disk.

61. A method for regenerating a disk mirror of claim 60 further comprising:

generating a spare disk by redistributing data from a third disk to a subset of the multiple of disks, after which the third disk can serve as the spare disk.

62. A method for regenerating a disk mirror of claim 61 wherein the step of redistributing data from the first disk further comprises reassigning blocks in a distribution map.

63. A method for regenerating a disk mirror in case of a processing unit failure in a system of multiple disks coupled to multiple processing units, said method comprising:

forwarding, by a first processing unit coupled to a disk containing a mirror of a first segment of a disk coupled with the failed processing unit, data from a secondary segment of the coupled disk to a spare processing unit;

writing, by the spare processing unit, data received from the first processing unit to a first segment of the disk coupled with the spare processing unit;

forwarding, by a second processing unit coupled to a disk containing a first segment mirrored by a secondary segment of the disk coupled to the failed processing unit, data from the first segment of the coupled disk to the spare processing unit; and

writing, by the spare processing unit, data data received from the second processing unit to a secondary segment of the disk coupled to the spare processing unit.